Rare Variant Association Testing in Case-Parent Trio Data

by

Linda Gai

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science

Baltimore, Maryland
April 2019

Abstract

The overall objective of this work is to develop a workflow for processing genome sequencing data from case-parent trios for rare variant analysis in the programming language R, and to provide publicly available software to allow others to perform rare variant analysis in case-parent trios themselves. In this thesis, we provide background on three different methods for rare variant detection designed for case-parent trios, detailing the analytical choices made in each method. We then illustrate our workflow by applying these methods to whole genome sequencing data from case-parent trios collected by the Gabriella Miller Kids First Initiative. All offspring in this dataset are affected by cleft-lip, with or without cleft palate (CL/P). We analyze both common and rare single nucleotide variants (SNVs) from the 8q24 region on chromosome 8q, a region that has been identified as containing a potential susceptibility locus for CL/P.

Advisors: Margaret Taub, Ph.D, Terri Beaty, Ph.D

Acknowledgements

I am deeply grateful to my thesis advisors, Dr. Margaret Taub and Dr. Terri Beaty, for their insight, patience, and support throughout this project and my time at the university generally. I would also like to thank my fellow students, Lacey Etzkorn and Xinye Li, for their support and help in the editing this thesis, and my family, for their love and encouragement.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Identifying the genetic causes of complex diseases is one of biomedicine's most urgently important goals. Genome-wide association studies (GWAS), which screen the genomes of hundreds or thousands of individuals in search of differences in allele frequencies between diseased and healthy subjects, represent the most common approaches to identifying disease-related genes; thousands have been conducted as of 2017 [1, 29]. However, variants identified by GWAS generally cause only mild increases in disease risk, leaving much of the heritability unexplained – even for diseases where large meta-analyses have been conducted [20]. One source of this "missing heritability" is thought to be rare gene variants conventional GWAS methods are too underpowered to detect, yet may exert large effects on risk [20].

In recent years, alternative study designs and statistical tests have been proposed to specifically address the problem of identifying causal rare variants, which classic GWAS approaches are largely unable to identify, because the allele frequency is too low to measure reliably. Family-based designs are one such approach: by including pedigree information, spurious associations caused by confounding due to and population stratification, which population-based study designs like GWAS are vulnerable to, can be largely avoided, thereby increasing statistical power. In particular, case-parent trio studies and their associated statistical methods were specifically designed to detect causal variants in admixed and otherwise heterogeneous populations [6]. The statistical power of tests for association between disease status and rare variants (i.e. variants with minor allele frequencies (MAF) <0.01) can also be increased further by grouping multiple variants together in a single test, instead of testing each variant one at a time. For example, the allelic transmission disequilibrium test (aTDT), and the genotypic transmission disequilibrium test (gTDT), both designed to assess whether a particular locus is either a disease susceptibility locus or in linkage disequilibrium (LD) with one, in case-parent trios, have recently been modified to test for rare variants by grouping together multiple variants in a combined or "collapsed" test. Both of these modified tests are referred to as the rare variant transmission disequilibrium test (RV-TDT [6] and rvTDT [15], respectively). Another way to aggregate multiple variants in one test is to apply a cluster-detection approach to identify small regions of the genome containing relatively high proportions of disease-associated variants; Scan-Trio is one method that does this [13].

In this work, we will examine these 3 methods for testing association between rare variants and disease status in trios: RV-TDT, rvTDT, and Scan-Trio. We begin by giving a detailed overview of the methods themselves, presenting them in the context of existing methods for analysis of common variants in case-parent trios. We then demonstrate a practical application of these methods by analyzing a dataset of 327

case-parent trios of European ancestry where the children were ascertained as being affected with orofacial cleft (i.e. cleft lip), with or without cleft palate (CL/P), from the Gabriella Miller Kids First (GMKF) Pediatric Research Program. Orofacial clefts, which include cleft lip only and cleft palate only, represent the most common group of craniofacial malformation in humans, affecting nearly 7,000 live births annually in the United States alone [31]. However, as of 2017, only about 20% of the estimated heritability can be explained using all recognized genetic risk factors identified through GWAS studies, suggesting that low frequency and rare variants might contribute significantly to the development of orofacial clefts [22]. Previous work has identified a region on chromosome 8, the gene desert 8q24, that appears to be strongly associated with CL/P in individuals of European ancestry, with peak signal located at 129.9-130.1 Mb on genome build hg18 [2, 22]. To better understand the genetic architecture underlying CL/P, and to determine whether any rare variants in the 8q24 region are causal, we applied these three approaches to rare variants in this region and compared the results.

Despite the promise of trio-based designs in detecting causal rare variants, there does not yet exist a standardized workflow for rare variant analysis in case-parent trios. Currently, trio data is analyzed with several different pieces of software, often in multiple programming languages, slowing methods development and limiting the ability of beginner and intermediate statisticians to perform trio-based analyses. As such, another major goal of this work is to develop a standardized workflow for using these three methods and integrate them into Bioconductor, a popular open-source platform for bioinformatics and biostatistics research with over 300,000 unique users annually [8], with the goal of creating a reproducible and easy-to-use workflow to facilitate analysis of and methods development for case-parent trio study designs.

2 Methods

2.1 Approaches to rare variant detection

To detect association between a particular genetic variant and a complex trait of interest, the classic GWAS approach is to test each variant individually for association. Association between a particular variant and the trait of interest is evaluated using a regression model: linear regression is generally used for continuous traits, while logistic regression is used for binary traits. However, while these single variant tests are widely-used, they are far less powerful for rare variants than for common variants simply due to their rarity, even if they have identical effect sizes [20]. Furthermore, because rare variants can differ greatly across different populations and ethnic groups, single variant tests frequently detect spurious rare variant associations in the presence of population substructure, further reducing power [6]. Thus, obtaining the sample sizes required

for a single variant test to detect causal rare variants is usually not feasible. To combat this, a few different approaches to increase the power of association methods have been developed. Here, we focus on two of them.

2.1.1 Burden tests

One approach is to aggregate multiple variants across a gene or region, thereby enriching the association signal and reducing the multiple testing penalty inherent in testing multiple variants at the same time. Generally, this approach is known as burden testing: instead of testing each variant individually, burden tests evaluate cumulative effects of multiple genetic variants within a gene or region, increasing power when multiple variants in the group are associated with a given disease and all have the same direction of effect (i.e., when the variants being tested are all deleterious, or all benign) [20]. Optionally, the power of these tests can be further increased by weighting each variant by some estimate of its probability of being causal: for example, each variant can be weighted by its frequency in the parents of affected offspring compared to the general population [15]; by its predicted functionality using functional annotation scores; or by its proximity to common variants associated with disease [13] (e.g., one could choose to examine only variants within a certain number of kilobases of a region that has high GWAS signal).

2.1.2 Using pedigree information

Because the transmission of alleles from parent to child should be unaffected by the ethnic makeup of the population at large, family-based test designs are not vulnerable to confounding due to admixture and population stratification. Although it is possible to reduce the effects of population substructure in population-based tests (e.g. case-control or cohort studies) by using principal component analysis (PCA), this approach does not work well in admixed populations even in tests for common variants [24]. Conversely, for family-based tests, the null and alternative hypotheses can be written in terms of the genotypes of the parents of affected offspring. By using Mendel's laws of allele transmission, the expected transmission probabilities of alleles or genotypes can be calculated exactly under the null hypothesis. The null hypothesis is a *composite* null hypothesis of both no linkage *and* no association between the genetic variant under examination and a possibly unobserved causal locus that affects disease risk: alleles in close proximity with one another are often transmitted together, so an allele that is transmitted to the offspring more often than is expected can be either a causal variant that directly affects disease risk, or can be a variant that is itself not directly causal, but is transmitted together with (possibly unobserved) causal variants [19]. Rejecting the null hypothesis, then, always implies the presence of both linkage between the variant and an unobserved causal locus and association between the tested variant and the trait of interest [18].

Table 1: **List of software packages for rare variant association tests in case-parent trios**

Method	Type	URL
RV-TDT	stand-alone	https://github.com/statgenetics/rv-tdt
rvTDT	R package (CRAN)	https://cran.r-project.org/web/packages/rvTDT/
Scan-Trio	R code	In development

2.1.3 Calculating significance via haplotype permutation

Many of the rare variant methods examined here use Monte Carlo simulations to assess statistical significance, because the distribution of the test statistic under the null hypothesis (i.e., that there is no association between child inheriting a particular variant and disease status in the child) is often difficult or impossible to derive: genetic variants, particularly those located in close proximity to one another, are often transmitted together, violating the assumption of independence required for many asymptotic results.

One solution is to permute the transmission status of haplotypes from parent to child. Under the null hypothesis, there is no association between rare variant transmission and disease status. Permuting the transmission status of each parental haplotype, then, will generate thousands of permuted case-parent trio datasets with "new" genotypes for each affected child. This allows us to obtain thousands of test statistics, all generated under the null hypothesis, while preserving the correlation in transmission status between variants that are frequently inherited together. The p-value of the observed test statistic is then the percentage of permuted statistics that are the same size or larger than the observed test statistic. For example, if we permute the dataset 1000 times, then we have 1000 permuted datasets, not including the observed dataset, and can therefore generate 1000 test statistics. If the test statistic calculated from the observed data is in the top 5% of the largest test statistics (i.e. it is in the top 50), then it reaches significance when $\alpha = 0.05$.

In the sections below, we introduce three recently-developed methods for detecting variants associated with disease risk in trios, all of which use the above approaches in tandem to increase statistical power: RV-TDT, rvTDT, and Scan-Trio. The homepages of these 3 methods are given in Table 1, and a summary is given in Table 2.

Table 2: **Comparison of rare-variant detection methods in case-parent trios**

Method	Description	Variants are weighted by...	Advantage	Disadvantage	Sub-tests
RV-TDT	detects whether a set of variants is transmitted more often from parents to affected offspring than would be expected by chance; uses McNemar's χ^2 test	the number of rare variants that heterozygous parents transmit to the cases	when risk of disease is related to number of rare variants transmitted from parents to offspring; when all variants have same direction of effect	unable to directly model risk of disease; limited power to detect mixture of deleterious and benign variants	TDT-CMC, TDT-BRV, TDT-WSS, TDT-VT
rvTDT	detects whether a genotype is transmitted more frequently from parents to affected offspring than would be expected by chance; based on conditional logistic regression for matched case-controls	the difference in genotype frequencies in the general population compared to parents of affected offspring	when disease risk is proportional to genotype's rarity in general population	obtaining population-level data for a matching population can be difficult	LC-1, LC-MAF, LC-PC; K-1, K-MAF, K-PC
Scan-Trio	identifies sets of disease-associated variants in close proximity with one another; based on cluster-detection scan statistic	N/A	when causal rare variants are in close proximity with one another; when all variants have same direction of effect	when causal rare variants are **not** in close proximity with one another; limited power to detect mixture of deleterious and benign variants	N/A

Abbreviations are as follows: BRV, Burden of Rare Variants; CMC, Combined Multivariate and Collapsing; K-1, Kernel Unweighted; LC-1, Linear Combination Unweighted; MAF, minor allele frequency; PC, population control; RV -TDT, Rare Variant Transmission Disequilibrium Test; rvTDT, rare variant Transmission Disequilibrium Test; VT, Variable Threshold; WSS, Weighted Sum Statistic

2.2 RV-TDT

The RV-TDT method extends the allelic transmission disequilibrium test (aTDT), which detects disease-associated common variants in case-parent trios, to the detection of rare variants. Specifically, the aTDT detects common *alleles* that are either disease-causing (or are in *linkage disequilibrium* (LD) with them) by comparing the *transmission* rate of major and minor alleles from each heterozygous parental locus (hence the name, "aTDT"). If, at such a locus, one allele is transmitted at a higher rate compared to the other, it is in both in LD and linked to an unobserved locus that is directly causal, since all of the offspring in the dataset are affected.

To improve the aTDT's ability to detect rare variants, the RV-TDT increases power by testing multiple variants in one test (e.g. by summing the number of transmitted alleles over a gene or individual), and weighting the variants according to one or more of the following methods:

1. the **Combined Multivariate and Collapsing (TDT-CMC)** statistic, which uses an indicator variable to denote whether **at least one** minor allele was transmitted from a parent to their affected offspring at any heterozygous locus

2. the **Burden of Rare Variants (TDT-BRV)** method, which uses the sum of indicator variables to denote the **total number** of minor alleles transmitted from a particular parent to their offspring at any heterozygous locus

3. the **Weighted Sum Statistic (TDT-WSS)**, which is similar to the BRV, but weights each variant site by the **inverse** of the variability in the number of times that its minor allele is transmitted from the parents to the affected offspring (i.e., if a minor allele at a particular locus or marker is almost always transmitted or almost never transmitted from a parent to their affected offspring, it probably has a bigger effect on risk and should be weighted more; but if it is transmitted to the affected offspring only some of the time, the minor allele at that variant site is probably not directly causal and should be weighted less)

4. the **Variable Threshold test (TDT-VT)**, which maximizes the test statistic (either CMC or BRV) over allele frequency cutoffs to define the set of variants included in each test region, and then finds an empirical p-value for the test statistic using permutation.

Difference in power between these 4 approaches is negligible, with the exception of the BRV and CMC, which are slightly more powerful than the other methods when the increased risk of disease from each additional inherited causal variants is constant, but is less powerful when the disease risk from inheriting causal variants is inversely related to the alleles' minor allele frequencies (MAFs) [7].

In the section below, we introduce the aTDT statistic and then develop in detail the 4 rare variant methods included in RV-TDT.

2.2.1 Allelic TDT

The aTDT, a straightforward application of McNemar's χ^2 test, assesses whether a genetic variant influences disease risk, or is in LD with a variant that influences disease risk. To do this, the aTDT compares major and minor allele transmission rates to affected offspring at heterozygous parental loci, under the null hypothesis of strict Mendelian transmission.

Consider a marker i with 2 possible alleles, M_1 and M_2. In a dataset containing n trios (and therefore $2n$ parents), the allele transmission events at marker i can be summarized in the following table:

	Non-transmitted allele		
Transmitted allele	M_1	M_2	**Total**
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

Intuitively, if marker i is completely independent of disease status (i.e., neither linked to an unobserved disease susceptibility locus, nor in linkage disequilibrium with one at a population-level), we would expect M_1 and M_2 to be transmitted from heterozygous parents to the affected offspring at the same rate as any other marker. Specifically,

$$\mathrm{P}(M_1 \text{ transmitted} \mid \text{parent is heterozygous}) = \mathrm{P}(M_2 \text{ transmitted} \mid \text{parent is heterozygous}) = 0.5.$$

Conversely, if i is either linked or in linkage disequilbirium with an unobserved disease susceptibility locus, we would expect that the disease risk variant would be transmitted at a higher rate in the affected offspring. (Note that the allele transmission events from homozygous parents (a and d) are uninformative – there is no randomness involved as the homozygous parents transmit the same allele no matter what.)

Since only heterozygous parents are informative, the number of heterozygous parents that transmit M_1 and not M_2 (denoted as b), and the number of heterozygous parents that transmit M_2 and not M_1 (denoted as c), contribute to the test statistic. If we define

$$p_b = \mathrm{P}(M_1 \text{ transmitted from heterozygous parents})$$

$$p_c = \mathrm{P}(M_2 \text{ transmitted from heterozygous parents}),$$

we have the null and alternative hypotheses

$$H_0 : p_b = p_c$$

$$H_1 : p_b \neq p_c$$

The McNemar's χ^2 test statistic is given by

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{1}$$

which is asymptotically χ^2-distributed with 1 degree of freedom.

The different variations on the RV-TDT depend on different ways of extending this test to accommodate multiple variants simultaneously. To accomodate this extension, in most cases, phasing the data and then assessing significance through permutations of the transmitted and untransmitted haplotypes is the preferred method. This will ensure proper control of Type I error and, in the case of the TDT-CMC, proper accounting for which combinations of alleles were transmitted together from each parent.

We start by considering an observed set of variants in our sample, denoted by L. Typically, these variants would be selected both based on their location (e.g., all variants inside a gene or other region of the genome) and their allele frequency (e.g., all variants with an MAF < 0.01 in the parents in our sample, or in an external database). Then the transmitted and unstransmitted alleles in each parent are examined and summarized as follows.

For each parent, we check each heterozygous variant site in variant set L and use an indicator variable to denote whether or not the minor allele was transmitted to the case.

Specifically, for every parent j at variant i in our variant set L, we define indicator variables c_{ij} and b_{ij} as

$$c_{ij} = \begin{cases} 1 & \text{if a minor-allele-transmitted event occurs for parent } j \text{ at variant } i, \\ 0 & \text{otherwise} \end{cases}$$

$$b_{ij} = \begin{cases} 1 & \text{if a major-allele-transmitted event occurs for parent } j \text{ at variant } i, \\ 0 & \text{otherwise} \end{cases}$$

Each variation of the RV-TDT uses a slightly different summary of these statistics as input into McNemar's test.

2.2.2 TDT-CMC

For both the TDT-CMC and the TDT-BRV, we start by counting the total number of transmitted minor and major alleles within each parent. That is, for parent j, we summarize the total number of transmitted minor alleles (c_j) and major alleles (b_j) as:

$$c_j = \sum_{i \in L} c_{ij}$$
$$b_j = \sum_{i \in L} b_{ij}$$

Note that $c_j + b_j$ gives the total number of heterozygous positions in parent j observed in variant set L.

The TDT-CMC tests for cumulative association between rare-variant-carrier status by combining data across parents based on the fraction of heterozygous sites where minor vs major alleles were transmitted from each parent.

The average percentage of transmitted alleles from the $2n$ parents that are **minor** alleles is proportional to (by a factor of $\frac{1}{2n}$)

$$c = \sum_{j=1}^{2n} c_j / (b_j + c_j)$$

and the average percentage of transmitted alleles from the $2n$ parents that are **major** alleles is proportional to

$$b = \sum_{j=1}^{2n} b_j / (b_j + c_j)$$

We then use these values to calculate the χ^2 statistic in Equation 1; a significant value indicates that we can reject the composite null hypothesis of no linkage or association between the collective rare variants in the genetic region and the disease. Note that because we separate the percentage of major and minor alleles transmitted by each parent j into the summands of b and c, respectively, every informative (i.e. heterozygous) parent contributes a score of 1 to the McNemar's test. To calculate the significance of the test statistic, we can either use the analytical solution (CMC-Analytical) or haplotype permutation (CMC-Haplo), though either way the haplotypes must be phased to ensure each parent's contribution to the McNemar's test is scored correctly and has a total weight of 1.

2.2.3 TDT-BRV

Instead of testing for a cumulative association between rare-carrier status and disease as in the TDT-CMC, the TDT-BRV tests the association between the number of rare variants transmitted (hence the name, "burden of rare variants" (BRV)) from each informative parent and disease status. In the TDT-BRV, each parent contributes a score equivalent to the number of heterozygous loci they carry (e.g. 1, 2, 3, ...), whereas, in the TDT-CMC, each informative parent can only contribute either a score of 1, decomposed into two fractions that indicate the proportions of heterozygous rare variants where major or minor alleles are transmitted, or 0 if they have no heterozygous rare variants in L.

Then for a data set with n trios (and $2n$ parents), the total number of transmitted minor alleles from all heterozygous parents c and total number of transmitted major alleles from all heterozygous parents b for gene or region L are given by

$$c = \sum_{j=1}^{2n} c_j$$
$$b = \sum_{j=1}^{2n} b_j$$

Again, these values c and b are used as input to McNemar's test (Equation 1). Significance is assessed using haplotype permutation (BRV-Haplo), to control Type I error in the presence of LD, so phasing is required for this test also.

2.2.4 TDT-WSS

In the TDT-WSS, the procedure is almost the same as the TDT-BRV, except each variant site is weighted by the inverse of the variability in the number of times the minor allele is transmitted to the affected offspring. By using this procedure, alleles that are almost always inherited, or almost never inherited, by the affected offspring are weighted more, while alleles that are inherited by the affected offspring only some of time are weighted less.

Specifically, for each variant i, let

q_i = minor allele frequency of variant i in the parental haplotypes **not** transmitted to the offspring.

Under the null hypothesis of equal probability of transmission of both alleles, the number of times the minor allele at variant site i is transmitted to the affected offspring can be thought of a Bin(n, q_i) random

variable, with an the estimated SD of

$$\widehat{\omega}_i = \sqrt{n \cdot q_i \left(1 - q_i\right)}.$$

Then, for parent j, we summarize the weighted number of transmitted minor alleles (c_j^{ω}) and major alleles (b_j^{ω}) as:

$$c_j^{\omega} = \sum_{i \in L} \frac{c_{ij}}{\widehat{\omega}_i}$$
$$b_j^{\omega} = \sum_{i \in L} \frac{b_{ij}}{\widehat{\omega}_i}$$

This gives c and b values of

$$c = \sum_{j=1}^{2n} c_j^{\omega}$$
$$b = \sum_{j=1}^{2n} b_j^{\omega}$$

Since the weights are calculated using estimates from the data, the p-values must be obtained empirically via permutation (WSS-Haplo) instead from the data directly, to avoid spurious associations.

2.2.5 TDT-VT

To increase the flexibility of how the set of markers, L, is defined, one can test multiple minor allele frequency (MAF) cut-offs to find the "optimal" MAF threshold that provides the highest level of significance for the test. This approach is known as variable threshold testing (hence the name "TDT-VT"). Statistical significance is assessed using haplotype permutation (VT-CMC-Haplo and VT-BRV-Haplo) as the test statistic is maximized over the various MAF allele frequencies and therefore must be corrected for multiple testing to ptoprtly control Type-I error.

2.2.6 Summary and limitations

In summary, the RV-TDT's main innovation is that it uses a variety of combining and weighting schemes, based on the number of rare variants heterozygous parents transmit to the cases, to improve the aTDT's ability to detect rare variants that may influence risk of a complex disease, while maintaining the aTDT's ability to minimize confounding due to population stratification. Furthermore, the 4 methods contained in the RV-TDT allow a researcher to use their prior knowledge to select the best method for a particular

scenario. Simple weighting schemes that count the number of rare alleles transmitted, like the TDT-BRV and TDT-CMC, are more powerful when the risk of disease is constant for the variants under examination, while more complex weighting schemes that take into account the rarity of the alleles, like the TDT-WSS or TDT-VT, are more powerful when the disease risk increases with the minor allele's rarity [15].

However, while the aTDT's intuitive design and ease of calculation make it an attractive choice, it is limited in several important ways that also extend to the RV-TDT. From a biological standpoint, the aTDT makes sometimes unrealistic assumptions about the relationship between disease risk and the transmitted marker alleles. For example, the aTDT implicitly assumes a multiplicative genetic model – that is, the risk of disease increases by a multiplicative factor for each additional risk allele at a particular susceptbility locus – because the parental contributions to disease risk (via transmission of alleles) are assumed to be independent [23]. As such, the aTDT does not perform as well when the underlying data-generating mechanism follows an alternative genetic model (e.g., dominant or recessive modes of inheritance), and the sample sizes required to achieve the same statistical power in these alternative models are often much larger (sometimes prohibitively large) [28]. (Modifications of the aTDT that assume these alternative genetic models exist, but are not commonly used [23].) Although evaluating RV-TDT's performance under different genetic models is beyond the scope of this work, it would be interesting to see whether RV-TDT also suffers from the same limitations.

Secondly, from a statistical standpoint, the aTDT is unable to directly model the relative risk of disease [27]. The aTDT, and by extension the RV-TDT, only reports p-values. One could imagine that a public health researcher would also be interested in estimates of the relative risk of disease for a particular set of rare variants, or in the confidence intervals for that estimate; RV-TDT cannot provide these estimates, only whether the set contains variants that are significant predictors of disease risk.

In the next section, we examine an alternative transmission disequilibrium test, the genotypic transmission disequilibrium test (gTDT) that detects linkage disequilibrium by comparing the transmission rate of possible *genotypes* at markers, rather than the transmission of possible alleles, to the case. We show that the gTDT addresses these sdhortcomings, at the cost of being more computationally intensive, and then show how the gTDT can be extended to rare variants, in a method known as rvTDT.

2.3 rvTDT

The rvTDT extends the genotypic transmission disequilibrium test (gTDT), which models the relative risk of disease comparing the child's genotype to the set of possible genotypes transmitted by the parents, to the detection of disease-causing rare variants. While the aTDT is an application of McNemar's chi-squared test, the gTDT is typically formulated as a the conditional logistic regression model for matched case-control

data [27, 4] to trio study designs. Here, the "cases" are the affected child in each trio, while the matched "controls" (or "pseudo-controls") are the other possible (but untransmitted) child genotypes, based on the parental genotypes at each locus where at least one parent is heterozygous. Each family therefore contributes one case and three pseudo-controls at each heterozygous locus. The gTDT is very flexible and, unlike the aTDT, can accommodate non-multiplicative genetic models (e.g. dominant, recessive, and co-dominant) [27].

In the rvTDT, we improve the gTDT's power to detect rare variants by weighting loci according to how common their variants are in the parental genotypes compared to the general population. If a heritable variant were in fact associated with disease, it should be found at higher rates in the parents of the affected offspring as well as in the affected offspring. Hence, transmitted alleles that are much more common in the parents compared to the general population are more likely to influence risk of disease and should be weighted more heavily. In addition, because the general population can be thought of as independent of the parents of the affected offspring in our dataset, estimating the weights of the variants using population data does not affect the asymptotic distributions of the test statistic [15], allowing us to make full use of the wealth of the available asymptotic results for the conditional logistic regression for matched-case control data.

In the sections below, we first develop the gTDT, describe the modifications needed to obtain the rvTDT, and then describe the rvTDT's benefits and shortcomings.

2.3.1 Genotypic TDT

The gTDT, also known as the genotype relative risk (gRR), examines each locus where at least 1 parent is heterozygous and models the probability that an affected child has its observed genotype, out of the set of possible child genotypes that heterozygous parents can produce. If a particular genotype is observed at a higher frequency in the affected offspring than would be expected by chance, it is plausible that the locus is associated with disease risk. For a data set consisting of n case-parent trios indexed by j, for $j = 1, \ldots, n$, let

$$A_j = \begin{cases} 1 & \text{if the child of trio } j \text{ is affected,} \\ 0 & \text{if the child of trio } j \text{ is unaffected} \end{cases}$$

$$c_{ij} = \text{ number of minor alleles observed at locus } i \text{ in the affected child of trio } j$$

$$x(c_{ij}) = \text{genetic effect of the offspring's genotype on disease risk}$$

For simplicity, we assume an additive genetic model, so that $x(c_{ij})$ is the number of minor alleles observed,

i.e. $x(c_{ij}) = c_{ij}$. Using logistic regression, we can model the relative risk of disease when the child of trio j has c_{ij} minor alleles at locus i, compared to the risk when the child has 0 minor alleles at locus i as

$$\log \frac{P(A_j = 1|C_{ij} = c_{ij})}{P(A_j = 1|C_{ij} = 0)} = \beta x(c_{ij}) \tag{2}$$

where β gives the additive increase in log-relative risk of disease in the affected child for genetic effect $x(c_{ij})$ at locus i, compared to $x(0)$. (Assuming an additive genetic model, β gives the increase in log-relative risk of disease when the affected child has c_{ij} minor alleles, compared to 0.) The probability that the child will have any of the possible genotypes under the assumption of Mendelian transmission is easily calculated if the parental genotypes are known. (For example, for a bilallelic locus i with major allele M_1 and minor allele M_2, 2 parents that are heterozygous at i can pass on one of 3 possible genotypes: M_1/M_1, M_1/M_2, M_2/M_1, with probability 1/4, 1/2, and 1/4, respectively.) Let us denote

$$p_{ij} = \text{parental genotypes for trio } j \text{ at locus } i$$
$$c_{ij} = \text{number of minor alleles in the affected offspring of trio } j \text{ at locus } i$$

Then for locus i, the probability that the child of trio j has the genotype c_{ij}, conditional on the child's affected status and the parental genotypes, is given by $P(C_{ij} = 1|A_j = 1, P_{ij})$. Using the laws of conditional probability, we end up with the conditional logistic regression model

$$\begin{aligned}
&P(C_{ij} = c_{ij}|A_j = 1, P_{ij}) \\
=&\frac{P(A_j = 1|C_{ij} = c_{ij}, P_{ij})P(C_{ij} = c_{ij}|P_{ij})}{\sum_{c'_{ij}=0}^{2} P(A_j = 1|C_{ij} = c'_{ij}, P_{ij})P(C_{ij} = c'_{ij}|P_{ij})} \\
=&\frac{\frac{P(A_j=1|C_{ij}=c_{ij})}{P(A_j=1|C_{ij}=0)} P(C_{ij} = c_{ij}|P_{ij})}{\sum_{c'_{ij}=0}^{2} \frac{P(A_j=1|C_{ij}=c'_{ij})}{P(A_j=1|C_{ij}=0)} P(C_{ij} = c'_{ij}|P_{ij})} \\
=&\frac{\exp\left[\beta \cdot x\left(c_{ij}\right)\right] P(C_{ij} = c_{ij}|P_{ij})}{\sum_{c'_{ij}=0}^{2} \exp\left[\beta \cdot x\left(c'_{ij}\right)\right] P(C_{ij} = c'_{ij}|P_{ij})} \text{ after plugging in (2).}
\end{aligned} \tag{3}$$

Note that the 2nd equality follows if we assume that, if we already know the affected offspring's number of minor alleles at a particular locus i, the parental genotypes at i do not give any additional information on the probability of the the child's disease status (i.e., $P(A_j = 1|C_{ij} = c_{ij}) = P(A_j = 1|C_{ij} = c_{ij}, P_j)$).

The statistical significance of β is assessed using a score test statistic. Note that $P(C_{ij} = c_{ij}|A_j = 1, P_{ij})$ is also the contribution from trio j to the likelihood $L_{ij}(\beta)$; by taking the log of (3) to find the log-likelihood, differentiating with respect to β, and evaluating at $\beta = 0$, we obtain the j^{th} trio's contribution to the score

test statistic for the i^{th} locus,

$$u_{ij} = x\left(c_{ij}\right) - \sum_{c'_{ij}=0}^{2} x\left(c'_{ij}\right) P(C_{ij} = c'_{ij} | P_{ij} = p_{ij})$$

The score test statistic for marker i is then given by

$$t_i = \frac{\left(\sum_{j=1}^{n} u_{ij}\right)^2}{\sum_{j=1}^{n} u_{ij}^2}$$

Under the null hypothesis of independence, the log-odds of risk of disease is unaffected by the child's genotype at this locus ($\beta = 0$) and t_i is asymptotically $\chi^2_{(1)}$ as $n \to \infty$.

2.3.2 Derivation of rvTDT

Like the modified aTDT used in RV-TDT, the gTDT's power is increased in rvTDT by testing multiple variants at once in a combined test. In this section, we introduce the 2 main weighting schemes used by rvTDT to increase statisticsl power.

Linear combination weighting A simple approach is to sum the score contributions across all variants i in a variant set L (e.g., a gene, for a gene-level test), weighted by the coefficient α_i, for each trio j by calculating

$$u_{\cdot j} = \sum_{i \in L} \alpha_i u_{ij}$$

and then summing the resulting score contributions over all n individuals. This gives us the linear combination (LC) test statistic:

$$t_{LC} = \frac{\left(\sum_{j=1}^{n} u_{\cdot j}\right)^2}{\sum_{j=1}^{n} u_{\cdot j}^2}$$

Like the score test statistic for the gTDT, t_{LC} of rvTDT is asymptotically $\chi^2_{(1)}$ distributed under the global null hypothesis (i.e. that none of the variants in the set L influence risk of disease).

Kernel weighting Another method is to weight the score contributions according to the contribution from each observed variant. To do this, we can sum the score contributions from a single locus across all n trios by calculating $u_{i\cdot} = \sum_{j=1}^{n} u_{ij}$, and then weight each by the coefficient α_i. We then square them, and

sum the squares across all variants within the gene. This gives us the kernel (K) test statistic:

$$t_K = \sum_{i \in L} (\alpha_i u_{i\cdot})^2$$

Note that this test statistic is a sum of quantities which may be correlated with one another, as score tests from nearby loci will be due to LD patterns in the data. However, it is possible to compute the covariance matrix of the set of single-locus score statistics, $U_{i\cdot}$. Then, under the global null hypothesis, it can be shown that t_K is a realization of a random variable T_K, where

$$T_K \sim \sum_{i \in L} \lambda_i \chi_1^2$$

where λ_i is the i^{th} eigenvalue of the covariance matrix of the scores across loci, $U_{i\cdot}$. Note that the test statistic is a linear combination of χ^2 variables; as such, we can use Davies method, which finds the distribution of linear combinations of χ^2 variables, to calculate the significance of t_K [15].

2.3.3 Weighting variants in the test statistic

Note that both statistics (t_{LC} and t_K) are functions of a vector of coefficients $\boldsymbol{\alpha}$, which represent some way of assigning importance or weights to all variants included in the test. There are a few different approaches to estimate the coefficients $\boldsymbol{\alpha}$ in the linear combinations of loci. The rvTDT R package estimates $\boldsymbol{\alpha}$ marginally (i.e., not taking the possibility of joint-effects into account), using 3 methods:

1. $\boldsymbol{\alpha_1}$, an unweighted estimate in which each variant under analysis is given equal weight (i.e. $\boldsymbol{\alpha} = \mathbf{1} = (1, \ldots, 1)^T$).

2. $\boldsymbol{\alpha_{MAF}}$, an estimate that inversely weights the variants by their rarity in the case-parent trios, using a weighting scheme used in the popular rare variant detection method, the Sequencing Kernel Association Test (SKAT) [9]. Specifically, each variant i is weighted by the probability of a Beta(1,25) to be equal to its MAF (i.e., for a random variable $X \sim Beta(1, 25)$, $\alpha_i = P(X = \text{MAF of variant } i)$).

3. $\boldsymbol{\alpha_{PC}}$, an estimate weighted by the difference in genotype frequencies between the parents and the general population, where PC denotes "population control".

$\boldsymbol{\alpha_1}$, and $\boldsymbol{\alpha_{MAF}}$ are primarily used as comparisons for the $\boldsymbol{\alpha_{PC}}$ approach, which is derived $\boldsymbol{\alpha_{PC}}$ in detail below.

Deriving $\boldsymbol{\alpha_{PC}}$: For the i^{th} locus, the estimate of $\alpha_{PC,i}$, $\tilde{\alpha}_{PC,i}$ is the Cochran-Armitage trend test statistic obtained from comparing the genotype frequencies of the parents of affected offspring (the "cases") at that locus, to those in the general population (the "controls"). The allele frequencies of the general population (required for the estimation of $\boldsymbol{\alpha_{PC}}$) can be obtained from very large, publicly available datasets. For example, the 1000 Genomes database contains the allele frequencies for several different populations and ethnicities, such as Europeans and sub-Saharan Africans [3]. We briefly introduce the Cochran-Armitage trend test statistic in the next section.

Cochran-Armitage trend test statistic: The Cochran-Armitage trend test statistic tests for an association between a variable with two categories (e.g., cases and controls) and an ordinal variable (e.g., low, medium, and high dosages). Here, we can think of the parents of the affected offspring as "cases" and the general population as controls. The dosage levels are the number of minor alleles in the genotype: 0, 1, or 2. For a locus i with 2 alleles, major allele M_1 and minor allele M_2, we denote these dosage levels as M_1/M_1, M_1/M_2, and M_2/M_2, respectively. The counts at each locus can be represented in tabular form below:

	M_1/M_1	M_1/M_2	M_2/M_2	**Sum**
Case (Parents of affected offspring)	N_{10}	N_{11}	N_{12}	$N_{1\bullet}$
Control (General population)	N_{00}	N_{01}	N_{02}	$N_{0\bullet}$
Column totals	$N_{\bullet 0}$	$N_{\bullet 1}$	$N_{\bullet 2}$	N

The trend test statistic is

$$\tilde{\alpha}_{PC,i} := \frac{N_{0\bullet} \times N_{1\bullet}}{N} \cdot \frac{\left[\sum_{k=0}^{2} k \times \left(\frac{N_{1k}}{N_{1\cdot}} - \frac{N_{0k}}{N_{0\cdot}}\right)\right]^2}{\left[\left(\sum_{k=0}^{2} k^2 \times \frac{N_{\bullet k}}{N}\right) - \left(\sum_{k=0}^{2} k \times \frac{N_{\bullet k}}{N}\right)^2\right]}$$

where k gives the weights of each genotype. The weights are chosen based on the hypothesized genetic model; assuming an additive model, $k = (0, 1, 2)$. The distribution of $\tilde{\alpha_j}$ is then asymptotically $\chi^2_{(1)}$ [21].

While the magnitude of $\tilde{\alpha}_{PC,i}$ is given by the Cochran-Armitage statistic, the sign of $\tilde{\alpha}_{PC,i}$ is given by the direction of the genotype frequency differences between the "cases" (the parents of the affected offspring) and "controls" (the population at large). When an allele at the i^{th} locus is more frequent in the parents than the general population, $\tilde{\alpha}_{PC,i}$ is positive, indicating that the variant may be associated with disease. Conversely, when the allele is less frequent in the parents, $\tilde{\alpha}_{PC,i}$ is negative, indicating that the variant may exert a protective effect.

2.3.4 Summary and limitations

In summary, rvTDT's main innovation is that it weights loci according to allele frequencies in the general population, creating a "data-driven" approach to weighting potential causal alleles. In comparison, RV-TDT uses weighting schemes derived only from the dataset under examination. In addition, because the rvTDT is based on the gTDT, for which many asymptotic and closed-form results have been derived, and because the population-based weighting scheme is independent of the dataset being analyzed, the asymptotic null distribution of the test statistic can be computed exactly, removing the need for the permutation tests of significance necessary in some RV-TDT tests.

The rvTDT also has downsides, owing to its reliance on population data and its more computationally demanding structure. Like the the RV-TDT, the rvTDT is designed to be robust to population stratification and admixture by accounting for family structure. However, if the population controls have extremely different ethnic makeup than the individuals in the dataset, differences in allele frequencies between the controls and the dataset could cause ethnicity-specific alleles to be overweighted, lowering power and potentially losing the main benefit of using a family-based association test in the first place [15]. In addition, if there is no population frequency information available for a variant, this variant will need to be excluded from the analysis. These effects can be mitigated by choosing a control population to match the dataset's population as closely as possible. For example, many large, public databases, such as 1000 Genomes, contain allele frequencies of several ethnic subpopulations [5]. Secondly, although we did not measure runtimes empirically, rvTDT takes considerably more time to run than RV-TDT – in our dataset, it took a few minutes to analyze the entire rare variants dataset, compared to the near-instantaneous run-time of RV-TDT.

In the next section, we discuss a third statistic, Scan-Trio, that uses the spatial correlation of disease-causing alleles to increase power.

2.4 Scan-Trio

Unlike RV-TDT and rvTDT, which are association tests that determine whether a *set* of rare variants is associated with risk of disease, Scan-Trio detects locations of genetic "hotspots" containing *clusters* of transmitted variants associated with risk of disease. The motivation for this approach is that disease-causing mutations tend to be located close together, in comparison to random but non-deleterious mutations [32]. Scan-Trio simultaneously tests all variants within a "window" that is either defined numerically (i.e., each window in the region contains the same number of markers) or spatially (i.e., each window contains all the markers within an kilobase interval of fixed size) for possible association with risk of disease. Consequently, the association signal is aggregated across all the variants in the window, thereby increasing the test's

statistical power to detect important rare variants, as well as common variants.

To test an entire genomic region, the window slides along the region of interest, and a new test statistic is calculated for each window [10]. The goal is to find a window – if there is one– that contains more transmitted alleles associated with disease than we'd expect to see by chance. Since all of the offspring in the dataset are affected, we would expect windows that contain disease-associated variants to have *more* transmitted minor alleles than expected by chance, and windows that don't contain disease-associated variants to have about the *same* number of transmitted minor alleles as we would expect to see by chance alone.

In the sections below, we derive the test statistic, describe how to calculate significance, and then discuss the major differences between the scan statistic approach and the other tests described previously.

2.4.1 Derivation

Let us first focus on testing for clustering in a single window W in a genomic region G. G contains n_G minor alleles across all parents, and $W \subseteq G$ contains a known number of minor alleles n_W, which may be enriched with disease-associated variants (usually minor alleles). For simplicity, we assume that all alleles are transmitted independently, we can describe the transmission of a minor allele at a particular position $i \in G$ from a parent to an affected offspring j as a Bernoulli random variable

$$c_{ij} = \begin{cases} 1 & \text{if a minor-allele-transmitted event occurs for parent } j \text{ at location } i \\ 0 & \text{otherwise} \end{cases}$$

where we allow the transmission rate to depend on whether the minor allele i is in the window W or not:

$$c_{ij} \sim Ber(p) \text{ for every position } i \in W \text{ which is heterozygous in parent } j$$
$$c_{ij} \sim Ber(q) \text{ for every position } i \notin W \text{ which is heterozygous in parent } j.$$

Then, define the total number of transmitted minor alleles in the region G from both parents to be $y_G = \sum_{j=1}^{2n} \sum_{i \in G} c_{ij}$, and define the total number of transmitted minor alleles in the window W as $y_W = \sum_{j=1}^{2n} \sum_{i \in W} c_{ij}$.

Again, since all the offspring in the dataset are cases, we can expect the transmission rate p of alleles within the window W to be greater than the transmission rate q of the alleles outside of the window. Accordingly, our null hypothesis is that no window contains a cluster of disease-associated variants, so $p = q$. The alternative hypothesis is that there is window containing a cluster of disease-associated variants, so $p > q$. Under H_1, a window of known size n_W contains a cluster of causal alleles, and the number of transmitted minor alleles in the window $y_W \sim Bin(n_W, p)$, and $y_G - y_W \sim Bin(n_G - n_W, q)$. Hence, the

likelihood under H_1 can be written as

$$L(p,q) \propto p^{y_W}(1-p)^{n_W-y_W} q^{y_G-y_W}(1-q)^{(n_G-n_W)-(y_G-y_W)}.$$

Under H_0 no window contains a cluster of disease-causing minor alleles, and $p = q$. Consequently under H_0, the likelihood for H_1 reduces to

$$L(p) \propto p^{y_G}(1-p)^{n_G-y_G}.$$

The maximum likelihood estimate under H_0 for p is then $\widehat{p_0} = \frac{y_G}{n_G}$. The estimates under the alternative hypothesis are $\widehat{p_1} = \frac{y_W}{n_W}$ and $\widehat{q_1} = \frac{y_G - y_W}{n_G - n_W}$. Taking the ratio of the likelihoods above and plugging in the maximum likelihood estimates for each of the parameters, we obtain the likelihood ratio statistic for window W:

$$LR_W = \begin{cases} \left(\frac{\widehat{p_1}}{\widehat{p_0}}\right)^{y_W} \left(\frac{1-\widehat{p_1}}{1-\widehat{p_0}}\right)^{n_W-y_W} \left(\frac{\widehat{q_1}}{\widehat{p_0}}\right)^{y_G-y_W} \left(\frac{1-\widehat{q_1}}{1-\widehat{p_0}}\right)^{n_G-n_W-(y_G-y_W)} & \widehat{p_1} > \widehat{q_1} \\ 1 & \text{otherwise.} \end{cases}$$

However, as we do not know the true location of the window, W, that harbors disease risk variants, we calculate the above statistic for a sequence of windows $w_k \subset G; k = 1 \ldots K$, along the region G.

First, we calculate the number of transmitted minor alleles in the sliding window $w_k \subset G$ is $y_k = \sum_{i,j} c_{ij}; i \in w_k$. The maximum likelihood estimates under the alternative hypothesis are now specific to the window: $\widehat{p_{1,k}} = \frac{y_k}{n_W}, \widehat{q_{1,k}} = \frac{y_G - y_k}{n_G - n_k}$. Hence, the likelihood ratio statistic for window w_k is

$$LR_k = \begin{cases} \left(\frac{\widehat{p_1}}{\widehat{p_0}}\right)^{y_k} \left(\frac{1-\widehat{p_1}}{1-\widehat{p_0}}\right)^{n_k-y_k} \left(\frac{\widehat{q_1}}{\widehat{p_0}}\right)^{y_G-y_k} \left(\frac{1-\widehat{q_1}}{1-\widehat{p_0}}\right)^{n_G-n_k-(y_G-y_k)} & \widehat{p_1} > \widehat{q_1} \\ 1 & \text{otherwise.} \end{cases}$$

2.4.2 Calculating significance

Because the transmission of variants within a small window is **not** independent, using the simplifying assumption of independence between transmissions is not reasonable, and we therefore cannot use results that rely on asymptotic distributions (e.g. χ^2 tests) to calculate the significance of the test statistic. Consequently, to obtain the distribution of the LR statistic and calculate its significance, we must use Monte Carlo simulation [17]. Specifically, we obtain replications of the dataset generated under the null hypothesis by permuting which haplotypes are transmitted from parent to child in the dataset many times (e.g. 1000 times), and

recalculating the maximum LR statistic across all the windows. By permuting the transmission of parental haplotypes, rather than the individual alleles, we preserve the spatial correlation caused by genetic linkage. The likelihood ratio statistic is calculated for each window, and the empirical p-value is the percentage of permuted LR statistics that are less than the LR statistics calculated from permuted data for that window. The cutoff for statistical significance is adjusted for multiple testing using a Bonferroni correction (i.e., an α-level of 0.05 would be divided by the number of windows K). Though this permutation procedure tests the null hypothesis of no association (i.e. $p = q = 0.5$), which is stronger than the null hypothesis that we want (i.e. $p = q$), it is a good approximation in practice, as previous work has shown the empirical Type I error is well-controlled in simulation experiments, even in the presence of disease-associated variants that do not cluster [11].

2.4.3 Summary and limitations

In summary, Scan-Trio's main innovation is that it uses the tendency of disease-causing variants to be located within close proximity to one other to increase power; instead of testing individual variants, Scan-Trio tests the entire set of markers located within a window of user-specified size. Additionally, unlike RV-TDT and rvTDT, Scan-Trio does not use any weighting schemes to increase power. Like RV-TDT, Scan-Trio relies only on the genotypes of the individuals under analysis to calculate the test statistic, and does not require any external information. This is in contrast to rvTDT, which requires genotype frequencies sampled from the general population.

Scan-Trio is best suited to situations where the researcher, *a priori*, has reason to believe that the risk variants are not uniformly distributed throughout the genomic region and are instead located within a relatively small section. For example, Scan-Trio might be well-suited to detecting rare variants in regions where an association between common variants and disease status has already been detected using more traditional methods (e.g., aTDT or gTDT). Generally, such analyses require rare variants to be filtered out, and it seems plausible that perhaps some of the "missing heritability" may be due to "losing" this information, especially since variants located close together tend to be transmitted together.

However, in cases where the rare variants are not clustered in a single set and are instead spread approximately evenly throughout the region, the power of Scan-Trio to detect disease-associated variants decreases substantially. In addition, here we assume, statistically, that the size of the window enriched with disease-causing variants is known. Realistically, however, we do not actually know the true window size, and are forced to test windows of various sizes. Worse still, Scan-Trio's power has been shown to decrease when the size of the scanning window is too large relative to the window containing the true risk-associated variants. The authors have proposed a "variable window approach" — i.e., we define the test statistic as the likelihood

ratio calculated with the window size that maximizes LR_W. The implementation of this method is beyond the scope of this project, but would surely improve the ease and efficacy of Scan-Trio.

3 Data

To compare the performance of these 3 methods, we analyzed whole genome sequence data from 327 case-parent trios of European ancestry from the Gabriella Miller Kids First (GMKF) Pediatric Research Program, identified because their child was affected by isolated, non-syndromic cleft lip, with or without cleft palate (CL/P). We focused specifically on a gene desert on chromosome 8, 8q24 (hg19 genome coordinates 117,700,001-146,364,022). A region of significant gTDT signal was found in 825 trios of European ancestry, suggesting that the 8q24 region may contain a gene or regulatory element tagged by SNPs involved in the development of CL/P [2, 22, 25]. Our goal in this work was to apply the workflow we developed to identify additional rare variants within 8q24, if there are any, and to compare the results between the different statistical methods discussed here.

3.1 Sample preparation and sequencing

Phenotypic data were collected by multiple research groups: University of Pittsburgh, University of Iowa, and the Johns Hopkins University. Individuals with birth defects other than CL/P (e.g. another structural malformation or major developmental delay) or evidence of a recognized Mendelian malformation syndrome were not included in this study. The research protocol and structured questionnaire was reviewed and approved by each recruitment site. Informed consent was provided by parents of the case and when appropriate assent from the case was also obtained recruitment.

Whole genome sequence (WGS) data was generated at the McDonnell Genome Institute (MGI) of Washington University, using reference genome hg19, consisting of 2x150bp paired end reads with approximately 30X coverage. An initial round of QC was performed at at MGI: variant call files with all samples genotyped at all positions were generated using GATK HaplotypeCaller; variant/sample QC was performed using Polymutt. Additional QC work was performed at the University of Pittsburgh.

3.2 Data cleaning

Multi-allelic variants were removed from analysis and only SNVs were considered (no indels). Genotype calls were filtered based on read depth and call quality: calls with read depth (DP) < 10 or call quality (GQ) < 20 were set to missing. Incomplete calls with one missing allele were set to missing and were phased.

3.3 Phasing

Prior to phasing, 15 trios that were outliers in the number of Mendelian inconsistencies ($> 100 errors$) were identified using PLINK software and were excluded from the dataset, leaving a total of 327 complete trios in the final dataset. Haplotype phasing was performed using BEAGLE 4.0 software [**?**]: in trios, BEAGLE infers the "transmitted" and "untransmitted" haplotypes for each parent, which is required for the permutation testing necessary to assess significance as described above. In addition, BEAGLE 4.0 imputes missing variant calls in trios using family information.

4 Analysis

4.1 Common variant analysis

First, to confirm whether previous evidence of linkage and association between common markers in 8q24 and CL/P is replicated in this dataset, and to compare performance between the trio methods for common variants and their rare variant extensions, an initial analysis was performed on the common variants in the dataset. SNVs with MAF < 0.01 were removed from the dataset, leaving a total of 3393 polymorphic SNVs. The aTDT and gTDT were performed using the *trio* R package from Bioconductor. Scan-Trio was also performed on this common variants dataset, to compare how it performs for both common variant and rare variant analysis. A window size of 100 markers, with overlap of 99, was arbitrarily chosen for this analysis, for a total of 16 windows. Significance was evaluated using 1,000 permutations. For an α-level of 0.05, the Bonferroni-corrected cutoff is $p = 0.003125$.

Minor alleles in a region of peak signal were strongly associated with increased risk of CL/P using all three methods (Fig. 1). Both the aTDT and gTDT yielded very similar results in the region of peak signal (hg19 chromosome 8:129.8-130 Mb). In the the interval 129,976,136-129,990,382, 18 loci were identified by the gTDT and 20 loci were identified by the aTDT as attaining p-values $< 10^{-6}$; of these, 1 was identified by the gTDT and 2 were identified by the aTDT as attaining a p-value of $< 10^{-8}$, which is the most widely used cutoff for genome-wide significance. Similarly, Scan-Trio identified 16 overlapping 100-marker windows located in the the interval 129,295,931-130,354,690 was significant ($p = 0.001$). These results replicate previous results about the 8q24 region [25], confirming the presence of some genetic element that influences risk of CL/P in individuals with European ancestry.

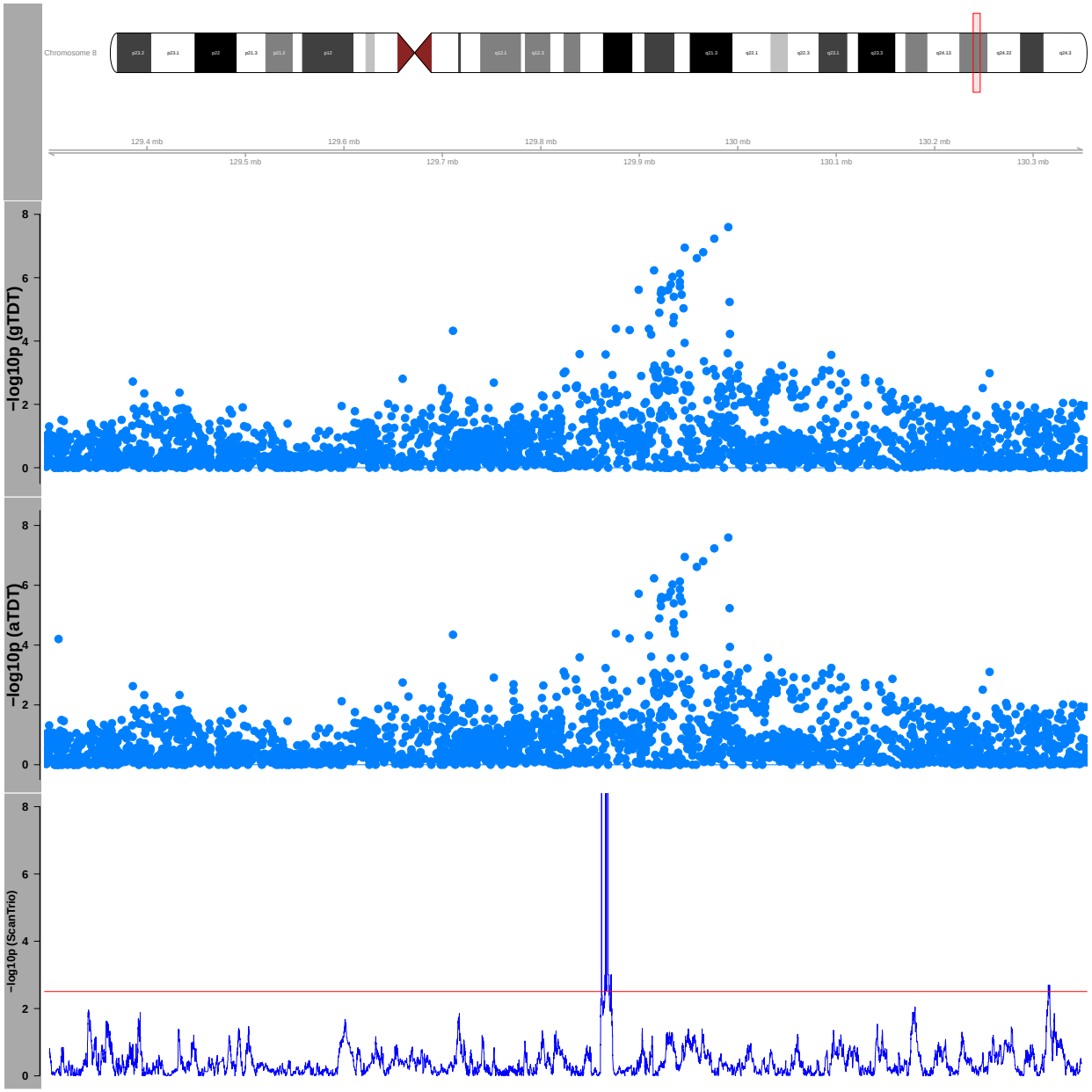

Figure 1: **aTDT, gTDT, and Scan-Trio results on common variants**. Graph of 8q24 region, and $-log_{10}(p)$ from analysis of common variants (MAF > 0.01). Red line indicates Bonferroni-corrected significance-level for $\alpha = 0.05$. X-axis shows genomic position (hg19). From top to bottom: i. Location of the 8q24 region. ii. aTDT results iii. gTDT results. iv. Scan-Trio results, for window size = 100 markers, 99 marker overlap.

4.2 Rare variant analysis

For the rare variant analysis, the dataset was restricted to SNVs with MAF < 0.01. Monomorphic SNVs were also removed, because they are non-informative. Subsequently, the remaining rare SNVs were filtered using annotation information to enrich rare variant signal, leaving a total of 336 SNVs, and analyzed using RV-TDT, rvTDT, and Scan-Trio.

4.2.1 Filtering using annotation information

To enrich rare-variant signal, the dataset was filtered by limiting the analysis to variants predicted to have causal effect using recommended score cut-offs for three popular measures of functional importance: EIGEN, GWAVA, and CADD. EIGEN is a measure of functional importance of genetic variants, both coding and noncoding, generated via an a weighted combination of other functional annotation measures [12]. GWAVA (Genome Wide Annotation of VAriants) is a measure of functional importance for non-coding variants, also generated by aggregating other functional measures; the authors used a threshold of > 0.5 to denote functional variants and ≤ 0.5 for non-functional variants [26]. CADD (Combined Annotation-Dependent Depletion) is a support vector machine-based functional annotation score that ranks variants by functionality and deleteriousness, based on annotation information; to identify pathenogenic variants, the developers recommend a cutoff between 10-20 [16]. As 8q24 is a non-coding region, we chose relatively lenient thresholds for filtering: SNVs with EIGEN scores > 4, GWAVA scores > 0.4 [26], or CADD scores > 10 were analyzed. All other SNVs were excluded from the dataset.

To ensure the datasets analyzed were the same for all 3 analytical methods, the dataset was further filtered to loci for which genotype frequencies in the general European population were available, as rvTDT requires estimates of genotype frequency in the general population. Genotype frequencies in the general European population for rvTDT were obtained from the 1000 Genomes (August 2015) European dataset [5]. All other annotation information, including functional annotation scores, was obtained from ANNOVAR (dated 02/01/2016) [30].

Finally, to take advantage of the tendency of disease-causing variants to be located close together, the rare variant set was broken down into consecutive, overlapping "windows" of 25 markers each, in a sliding window-based approach. A total of 344 25-marker windows were produced. RV-TDT, rvTDT, and Scan-Trio were then applied to each window, and the significance cut-off was Bonferroni corrected (i.e, for 344 windows, $\alpha = 0.05/344 = 0.00014$).

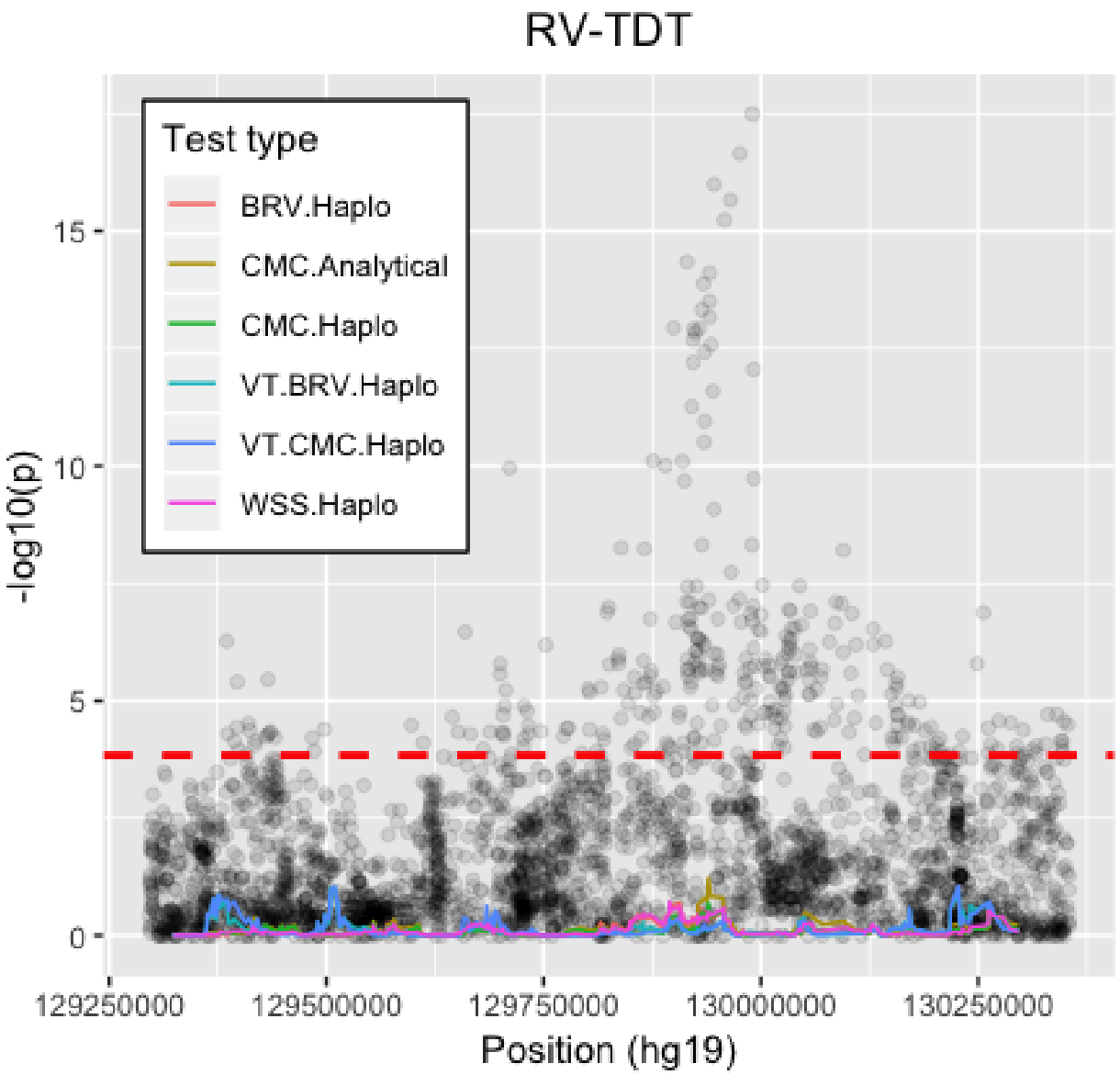

Figure 2: **RV-TDT results on rare variants, compared to gTDT results on common variants**. Graph of RV-TDT $-\log_{10}(p)$-values from rare variant (MAF < 0.01) analysis, using scanning windows of 25 markers, with 24 marker overlap. Gray circles represent results of gTDT using polymorphic, common variants (MAF>0.01). X-coordinate denotes genomic position (hg19) at center of window. Red dotted line indicates Bonferroni-corrected significance-level for $\alpha = 0.05$.

4.2.2 RV-TDT

We applied six variants of the RV-TDT: BRV, two types of CMC (with p-values calculated analytically and empirically, using permutations), VT-BRV, VT-CMC and WSS. For the VT methods, MAFs < 0.01 were used for marker inclusion. None of the RV-TDT sub-tests came remotely close to reaching statistical significance (Fig. 2), though p-values differed based on the analytical scheme used.

In general, the most significant p-values were produced by methods that simply indicated whether the parents transmitted a rare variant or not (CMC-Analytical, CMC-Haplo, VT-CMC-Haplo), while methods with the highest p-values were weighted by the number of variants transmitted per parent (BRV-Haplo, VT-BRV-Haplo, and WSS-Haplo). The CMC-Analytical method had relatively significant p-values located close to the gTDT peak, as did WSS-Haplo (p = 0.1-0.3). Conversely, BRV-Haplo, VT-BRV-Haplo, and WSS-Haplo gave relatively higher p-values close to the gTDT peak (0.5-0.75). BRV counts the number of rare variants transmitted, WSS weights each variant site by the number of variants in the parental haplotypes that are not transmitted in the parental offspring, and VT maximizes the test statistic over the allele frequencies. This result potentially suggests a very small number of rare variants, if any, are associated with disease in the region.

4.2.3 rvTDT

As with RV-TDT, none of rvTDT's subtests came close to reaching significance, though they came slightly closer to reaching significance than RV-TDT (Fig. 3). Overall, the p-values from the sub-tests using linear combinations to weight the loci tended to be more significant than those using kernels, with the linear combinations weighted by the population controls giving the most significant p-values of all. Specifically, the most significant p-values were reported when the weights for each locus were modeled using a linear combination, weighted by their allele frequencies in the general population (LC-PC) and using a linear combination weighted by the inverse of the minor allele frequencies (LC-MAF), which yielded nearly identical p-values. Overall, the locations of the peaks in the RV-TDT results were not at all similar to the locations of the peaks in the common variants analysis.

4.2.4 Scan-Trio

As with the previous 2 methods, no windows produced a significant result (Fig. 4). The p-value for every window in the region was 1, suggesting that each window either contains very few or 0 rare variants that influence risk of CL/P.

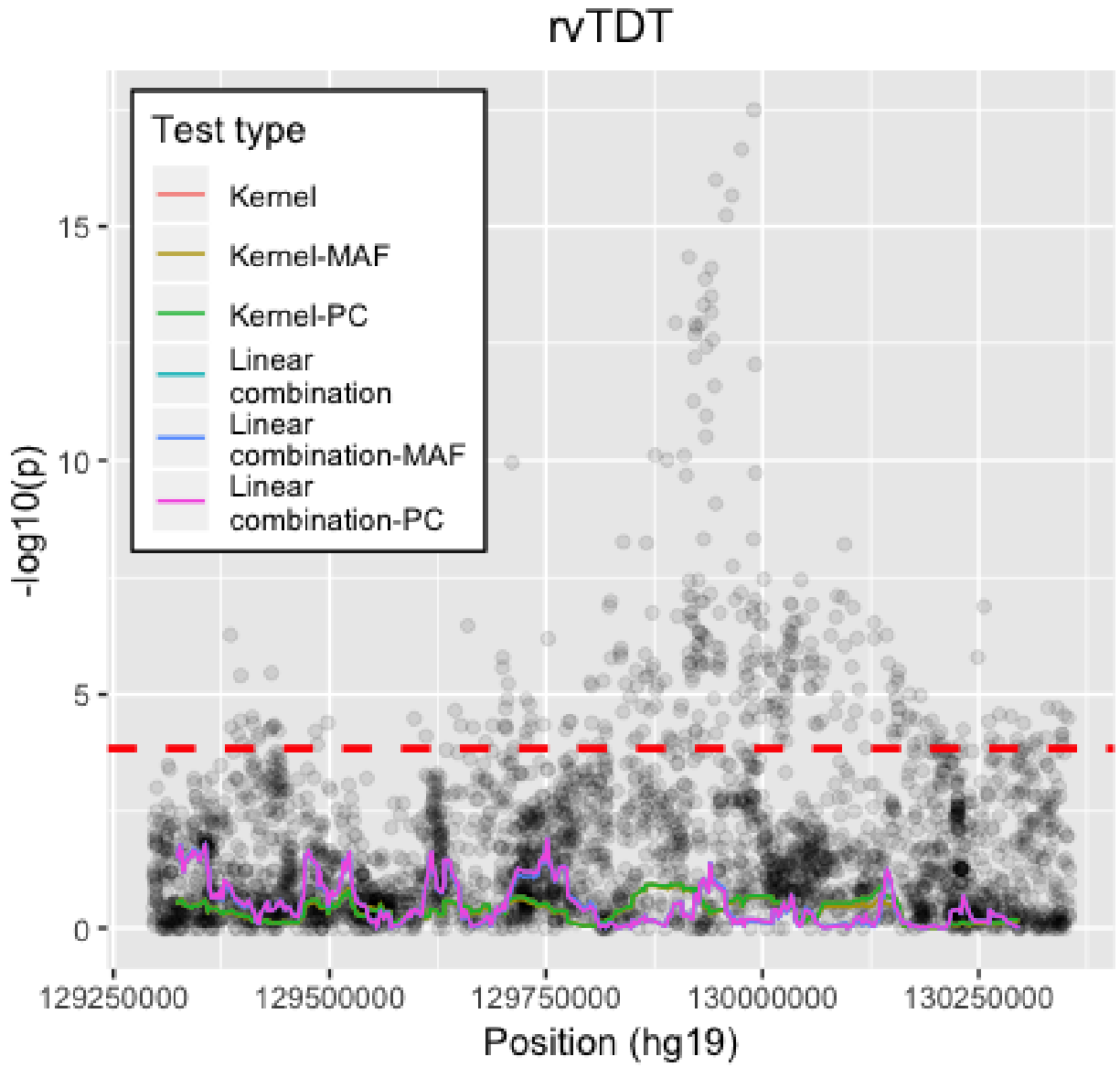

Figure 3: **rvTDT results on rare variants, compared to gTDT results on common variants**. Graph of rvTDT $-\log_{10}(p)$-values from rare variants (MAF > 0.01) analysis, using scanning windows of 25 markers, with 24 marker overlap between windows. Gray circles represent results of gTDT using polymorphic, common variants (MAF<0.01). X-coordinate denotes genomic position (hg19) at center of each window. Red dotted line indicates Bonferroni-corrected significance-level for $\alpha = 0.05$.

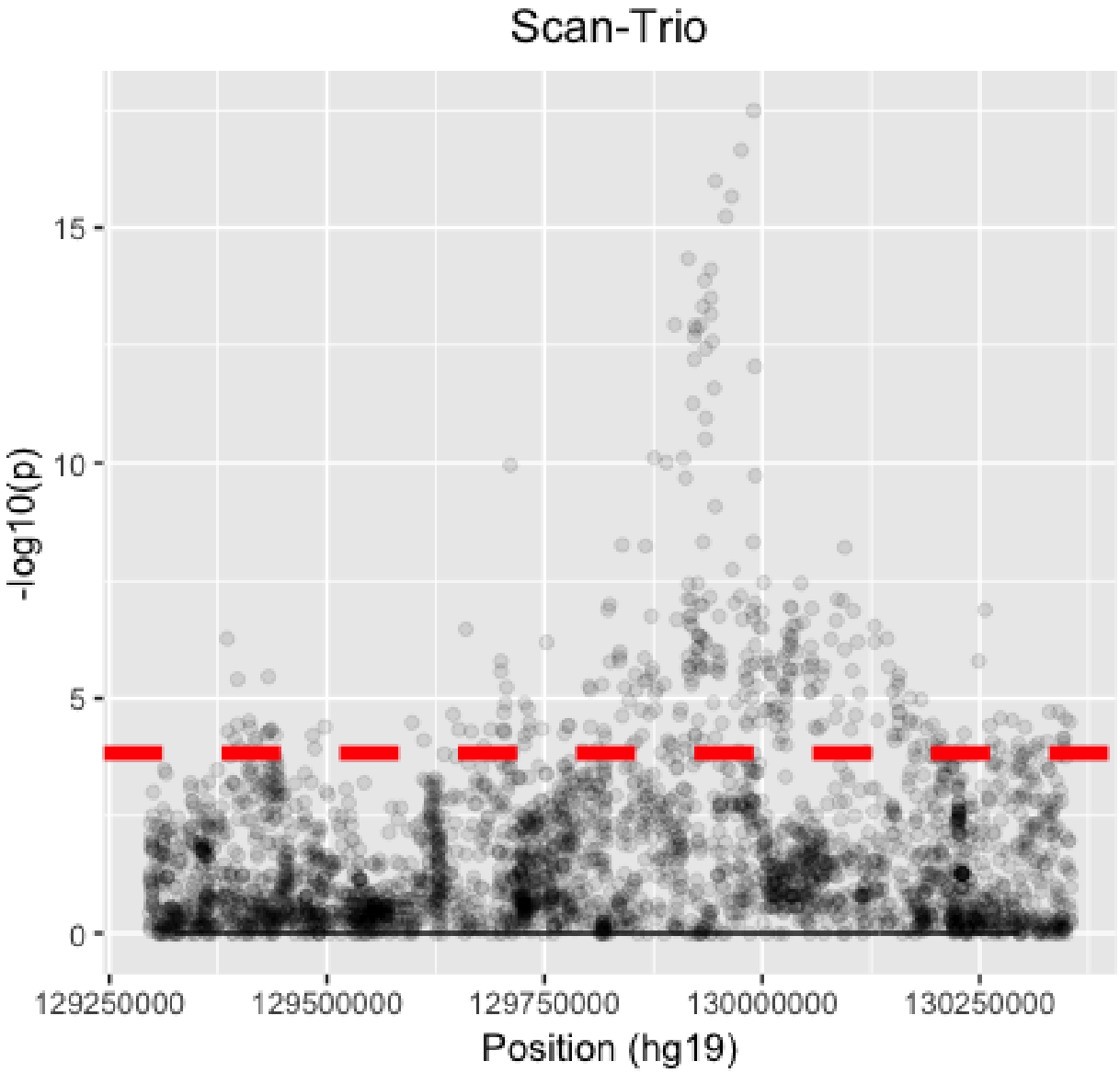

Figure 4: **Scan-Trio results on rare variants, compared to gTDT results on common variants**. Graph of Scan-Trio -$\log_{10}(p)$-values from rare variants (MAF < 0.01) analysis, using scanning windows of 25 markers, with 24 marker overlap between windows. Gray circles represent results of gTDT using polymorphic, common variants (MAF>0.01). X-coordinate denotes genomic position (hg19) at center of each window. Red dotted line indicates Bonferroni-corrected significance-level for $\alpha = 0.05$.

4.3 Summary

Although 129.8-130.3 Mb (hg19) in the 8q24 region was strongly associated with risk of CL/P in common variants using all 3 common-variant detection methods examined in this work, the rare variants in this dataset did not reach significance using any rare-variant extension of the common variants tests (Fig. 5). As such, we did not find evidence to suggest that rare variants in the 8q24 region influence risk of CL/P.

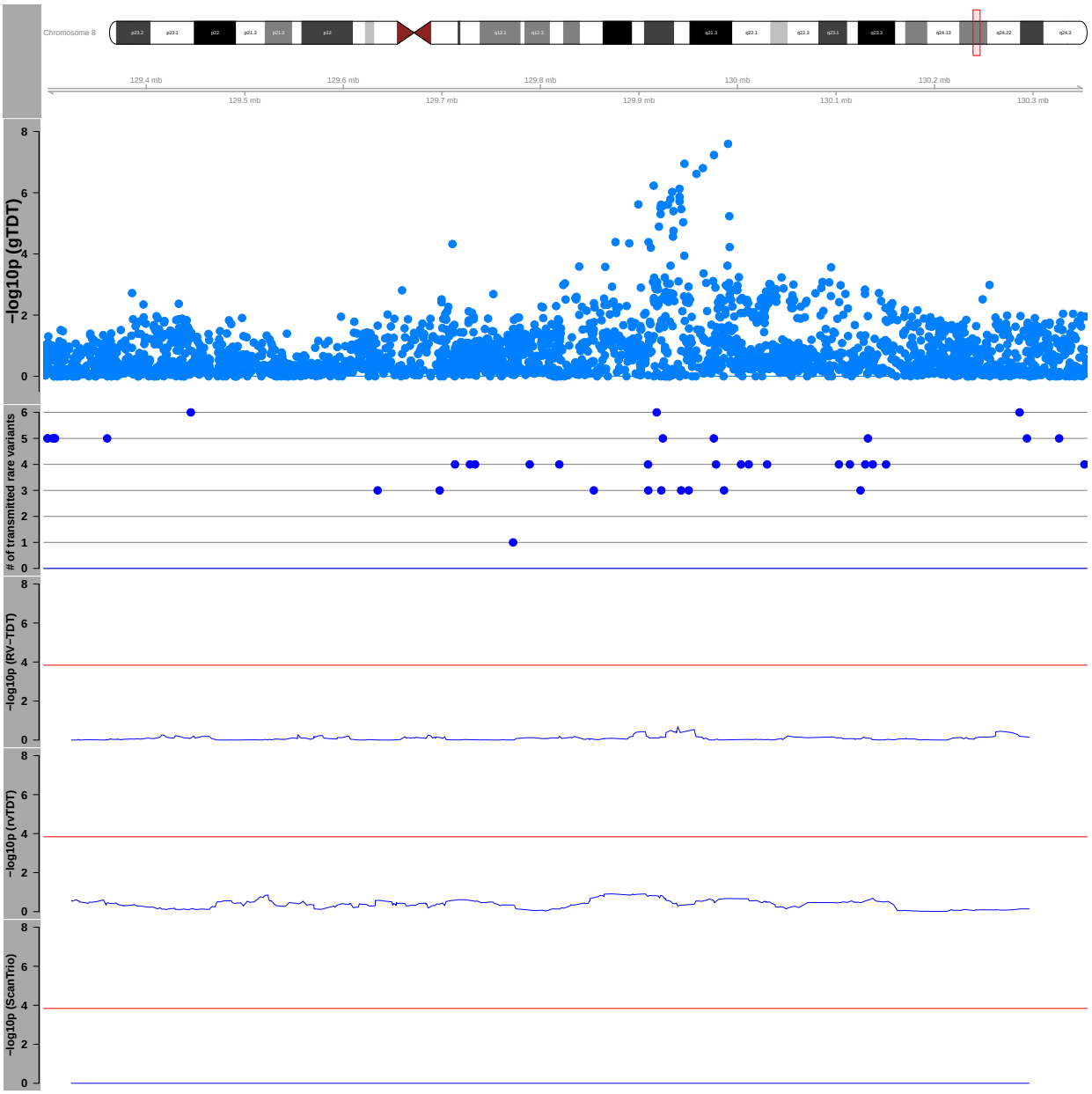

Figure 5: **Comparison of results between rare variant and common variants analyses**. Graph of 8q24 region, transmitted rare variant counts, and $-log_{10}(p)$ from gTDT performed on common (MAF $>$ 0.01) variants, and 3 methods to detect the combined effects of rare variants (MAF $<$ 0.01) on disease risk, using windows of size = 25 markers, 24 marker overlap. Red line indicates Bonferroni-corrected significance-level for $\alpha = 0.05$. X-axis shows genomic position (hg19). From top to bottom: i. Location of 8q24 region on chromosome 8. ii. gTDT results for common variants. iii. Transmitted rare variant counts. iv. RV-TDT results for rare variants predicted to be deleterious. v. rvTDT results for the same rare variants. vi. Scan-Trio results for the same rare variants.

5 Software

A primary goal of this work was to develop an effective and efficient case-parent trio analysis workflow in R, using packages from CRAN and Bioconductor, that will easily allow researchers to replicate the above analysis on their own datasets and facilitate methods development for case-parent trios. As such, code for this analysis and a workflow (in-progress) are available on Github: `https://github.com/lindagai/8q24_project/tree/master/code2`. The code is divided into 3 main sections:

1. **Data cleaning**

   A VCF file of variant calls from the whole genome sequencing data are cleaned and phased using BEAGLE 4.0. The phenotype file describing the familial relationships of all study participants is edited to ensure the names of the individuals match those in the VCF. 2 datasets, one containing common variants and the other containing rare variants, are created. Annotation information from ANNOVAR is used to filter the rare variants based on user-specified filters, to limit the rare variant analysis to SNVs predicted to be deleterious.

2. **Methods**

   Common variants analysis is performed using *trio*, using the aTDT and gTDT, as well as Scan-Trio. In addition, code is provided for breaking the rare variant dataset into windows containing a number of markers specified by the user, running the 3 rare variants methods on each of the windows, extracting the results from each window, and assembling them into 1 table per method.

3. **Visualization**

   Individual Manhattan plots for all 4 methods described in this workflow are created using *ggplot*2. Manhattan plots comparing the aTDT, gTDT, and Scan-Trio results from the common variants analysis, and Manhattan plots to compare the gTDT results from the common variants and the RV-TDT, rvTDT, and Scan-Trio results from the rare variants analysis were created using $GViz$. Code for calculating the number of rare variants transmitted to the offspring is also provided.

6 Discussion

In summary, we have presented here a suite of 3 statistical methods for detecting combined effects of rare variants in case-parent trios, and present a workflow and example analysis that will allow statistical geneticists to perform common and rare variants analysis in their own sequencing data from case-parent trios. We explain how to derive these 3 methods in detail, what assumptions they make, and when to use each one

(Table 1). Although we did not find positive results in the dataset of European case-parent trios analyzed here, our intention is to use this workflow to analyze other regions of the genome for rare variants associated with risk of CL/P. In this section, we will describe the contributions of this work and their limitations, as well as directions for future work.

We focused here on giving an overview of 3 recently-developed methods for rare variant detection, deriving each in detail, and describing their respective strengths and limitations. To our knowledge, as of 2019, no broad review of statistical methods in rare variant detection in case-parent trios has yet been done. Each of the methods presented here takes a different approach to aggregating rare variant signals and testing for possible effects on risk to a complex heterogenous disease like CL/P, and each may be appropriate for different settings, based on how the specific rare variants under examination are related to disease risk. However, because the simulation conditions used to evaluate each method were different (as they were performed by different groups of researchers), differences in performance between these 3 methods are difficult to evaluate accurately. To allow researchers to better interpret their results from using these methods, a simulation study directly comparing the performance of these 3 methods under different conditions should be performed.

A second major goal of this project was to bring common and rare variant analyses for case-parent trios up to the standards of reproducibility demanded by modern science. We provide a workflow that will allow a user to run a rare variants analysis, using cutting-edge techniques, from start-to-finish in R. This both streamlines methods development for case-parent trios by making trio analysis simpler to perform, as well as reduces the need for researchers to perform hard-to-reproduce, piecemeal trio analyses using many different pieces of software.

However, further work is required to make this workflow easier to use. For example, in this analysis, we assume the windows enriched with disease-associated variants can be detected with sliding windows of size 25 variants, though of course we do not know the actual size of the window. Extending the workflow to allow multiple window sizes to be tested and graphed simultaneously would make the package more user-friendly. In addition, we hope to implement functionality to test windows defined spatially (e.g., all markers within a kilo-base interval of length specified by the user); currently, we only allow the user to specify the number of variants in each window.

Extensions of some of these methods examined here may be useful to incorporate into the workflow in the future. For example, the authors of rvTDT have recently derived an approach to estimate the weights of each variant jointly at all variant sites simultaneously, using individual-level genotype data for both the data set being analyzed and the general population [15, 14]. This could be done by fitting a logistic regression model to all the informative variant sites, including individual genotypes as predictor variables in a multiple regression model, and using the parent or general population status as the case-control outcomes.

Under simulation, joint estimation of the weights of each variant results in models that perform better than those produced by using the marginal estimation approach described earlier in the paper [15]. However, joint estimation requires genetic data for each individual at each variant site, and although individual-level genetic sequence data is available in some public datasets for some genetic regions, the datasets are often much smaller. As such, gains in performance are often undercut by the availability and quality of data on the general population. As public genome resources grow, however, individual-level genetic sequence data should become more widely available, and users would benefit from the option to jointly estimate the optimal weights of each locus in rvTDT if population data is available. In addition, the authors of Scan-Trio have proposed a variable-window approach to systematically identify the "optimal" window size that maximizes the likelihood ratio statistic LR_W [13], as opposed to examining window sizes one at a time. Though implementing this method was beyond the scope of this project, implementing this approach would be helpful for analysts who do not have *a priori* knowledge of the size of the window they want to examine.

The third and last major component of this project was to analyze a recognized region of the genome thought to contain an element (coding or regulatory) that influences risk of CL/P. Though we found that no groups of rare variants reached statistical significance using any of the 3 methods used here, we hope to use this workflow to analyze additional regions of the genome reported to harbor causal genes or other regulatory elements that may affect disease risk for CL/P. Moreover, the analysis presented here was limited in several ways. For example, variant calling and phasing potentially introduces additional problems into the analysis. First, the variant-calling software (GATK Haplotype Caller) used for this analysis did not explicitly take family structure into account. This can lead to inconsistent calls between family members, potentially introducing systematic errors that can lead to bias in genetic association test statistics. Secondly, we chose to impute half-calls and missing observations by using the phasing software BEAGLE 4.0. For some of the methods examined, phasing can inflate type I error. For example, in RV-TDT, phasing can increase false-positive rates, so the authors instead chose to remove the imputed variant calls from their analysis [6]. If conservative estimates of disease risk from rare variants are desired, the best course of action may be to instead remove the half-missing and missing variant calls from analysis.

Furthermore, relatively few annotation resources are available for the 8q24 region, as it is a gene desert. The dearth of annotation resources may have especially impacted the rvTDT results, as it relies on information like genotype counts in the general population to estimate variant weights. However, population genotype counts in this analysis were estimated using allele frequency, as genotype counts were not available, though even allele frequencies were only available for relatively few loci (2021 out of 3393 SNVs, prior to filtering). An additional layer of filtering by functional annotation scores also drastically reduced the number of SNVs available for analysis, to 368 from 2021, even though the filtering cutoffs were chosen to be lenient.

Some of these problems can be addressed by simply analyzing other regions for which more extensive annotation information is available, such as coding regions. For non-coding regions like 8q24, however, comparing results between tests before and after filtering by annotation information may be necessary, as filtering using annotation information greatly reduces the number of SNVs available for analysis.

Disease risk variants too rare to be detected by population-based GWAS methods are hypothesized to be one of the root causes of the missing heritability problem, and in many common genetic diseases, much heritability remains to be explained. Because the case-parent trio design was expressly created to reduce confounding from population stratification, it has considerable advantages in detecting rare and low frequency variants that may influence disease risk. In this work, we have provided several resources to facilitate rare variant analysis of case-parent trio data: 1) a review of state-of-the-art analytical methods to detect rare variants associated with increased disease risk in case-parent trio sequencing data; 2) software that allows researchers to easily and efficiently apply these newly-developed methods on their own sequencing data from case-parent trios, while meeting contemporary standards of reproducibility; and 3) an example analysis that beginner and intermediate statistical geneticists can apply to their own sequencing data from case-parent trios, where we searched for risk variants in a genetic region linked to cleft lip with or without cleft palate, a common genetic disease for which much of the heritability is still unexplained. By enabling scientists to search for additional rare disease risk variants in case-parent trio studies, we hope this work will facilitate the development of additional methods for rare variant analysis for case-parent trios designs, and expedite the search for missing heritability in complex, heritable disease in general.

7 Bibliography

References

[1] Baishali Bandyopadhyay, Veda Chanda, and Yupeng Wang. Finding the sources of missing heritability within rare variants through simulation. *Bioinformatics and biology insights*, 11:1177932217735096, 2017.

[2] Terri H Beaty, Jeffrey C Murray, Mary L Marazita, Ronald G Munger, Ingo Ruczinski, Jacqueline B Hetmanski, Kung Yee Liang, Tao Wu, Tanda Murray, M Daniele Fallin, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near mafb and abca4. *Nature genetics*, 42(6):525, 2010.

[3] Ewan Birney and Nicole Soranzo. Human genomics: The end of the start for population sequencing. *Nature*, 526(7571):52, 2015.

[4] NE Breslow, NE Day, KT Halvorsen, RL Prentice, and C Sabai. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108(4):299–307, 1978.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[6] Zongxiao He, Brian J O'Roak, Joshua D Smith, Gao Wang, Stanley Hooker, Regie Lyn P Santos-Cortez, Biao Li, Mengyuan Kan, Nik Krumm, Deborah A Nickerson, et al. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *The American Journal of Human Genetics*, 94(1):33–46, 2014.

[7] Zongxiao He, Di Zhang, Alan E Renton, Biao Li, Linhai Zhao, Gao T Wang, Alison M Goate, Richard Mayeux, and Suzanne M Leal. The rare-variant generalized disequilibrium test for association analysis of nuclear and extended pedigrees with application to alzheimer disease wgs data. *The American Journal of Human Genetics*, 100(2):193–204, 2017.

[8] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115, 2015.

[9] Iuliana Ionita-Laza, Seunggeun Lee, Vladimir Makarov, Joseph D Buxbaum, and Xihong Lin. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21(10):1158, 2013.

[10] Iuliana Ionita-Laza, Vlad Makarov, Joseph D Buxbaum, ARRA Autism Sequencing Consortium, et al. Scan-statistic approach identifies clusters of rare disease variants in lrp2, a gene linked and associated with autism spectrum disorders, in three datasets. *The American Journal of Human Genetics*, 90(6):1002–1013, 2012.

[11] Iuliana Ionita-Laza, Vlad Makarov, Joseph D Buxbaum, ARRA Autism Sequencing Consortium, et al. Supplementary information: Scan-statistic approach identifies clusters of rare disease variants in lrp2, a gene linked and associated with autism spectrum disorders, in three datasets. *The American Journal of Human Genetics*, 90(6):1002–1013, 2012.

[12] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 48(2):214, 2016.

[13] Iuliana Ionita-Laza, Bin Xu, Vlad Makarov, Joseph D Buxbaum, J Louw Roos, Joseph A Gogos, and Maria Karayiorgou. Scan statistic-based analysis of exome sequencing data identifies fan1 at 15q13. 3 as a susceptibility gene for schizophrenia and autism. *Proceedings of the National Academy of Sciences*, 111(1):343–348, 2014.

[14] Yu Jiang, Yunqi Ji, Alexander B Sibley, Yi-Ju Li, and Andrew S Allen. Leveraging population information in family-based rare variant association analyses of quantitative traits. *Genetic epidemiology*, 41(2):98–107, 2017.

[15] Yu Jiang, Glen A Satten, Yujun Han, Michael P Epstein, Erin L Heinzen, David B Goldstein, and Andrew S Allen. Utilizing population controls in rare-variant case-parent association tests. *The American Journal of Human Genetics*, 94(6):845–853, 2014.

[16] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[17] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[18] Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385, 2006.

[19] Nan M Laird and Christoph Lange. Family-based methods for linkage and association analysis. *Advances in genetics*, 60:219–252, 2008.

[20] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014.

[21] Wen-Chung Lee. Optimal trend tests for genetic association studies of heterogeneous diseases. *Scientific reports*, 6:27821, 2016.

[22] Elizabeth J Leslie, Jenna C Carlson, John R Shaffer, Carmen J Buxó, Eduardo E Castilla, Kaare Christensen, Frederic WB Deleyiannis, Leigh L Field, Jacqueline T Hecht, Lina Moreno, et al. Association

studies of low-frequency coding variants in nonsyndromic cleft lip with or without cleft palate. *American Journal of Medical Genetics Part A*, 173(6):1531–1538, 2017.

[23] Cathryn M Lewis. Genetic association studies: design, analysis and interpretation. *Briefings in bioinformatics*, 3(2):146–153, 2002.

[24] Jinghua Liu, Juan Pablo Lewinger, Frank D Gilliland, W James Gauderman, and David V Conti. Confounding and heterogeneity in genetic association studies with admixed populations. *American journal of epidemiology*, 177(4):351–360, 2013.

[25] Tanda Murray, Margaret A Taub, Ingo Ruczinski, Alan F Scott, Jacqueline B Hetmanski, Holger Schwender, Poorav Patel, Tian Xiao Zhang, Ronald G Munger, Allen J Wilcox, et al. Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between a sians and e uropeans. *Genetic epidemiology*, 36(4):392–399, 2012.

[26] Graham RS Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294, 2014.

[27] Holger Schwender, Margaret A Taub, Terri H Beaty, Mary L Marazita, and Ingo Ruczinski. Rapid testing of snps and gene–environment interactions in case–parent trio data based on exact analytic parameter estimation. *Biometrics*, 68(3):766–773, 2012.

[28] Susan L Slager, J Huang, and VJ Vieland. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 18(2):143–156, 2000.

[29] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[30] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[31] Mahsa M Yazdy, Margaret A Honein, Sonja A Rasmussen, and Jaime L Frias. Priorities for future public health research in orofacial clefts. *The Cleft palate-craniofacial journal*, 44(4):351–357, 2007.

[32] Peng Yue, William F Forrest, Joshua S Kaminker, Scott Lohr, Zemin Zhang, and Guy Cavet. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human mutation*, 31(3):264–271, 2010.

Linda Gai

1913 Bank St
Baltimore, MD 21231
lindagai@jhu.edu

Johns Hopkins University

Education

Johns Hopkins University - Baltimore, MD GPA 3.7
- ScM candidate in Biostatistics, expected graduation Spring 2019

Swarthmore College - Swarthmore, PA
- B.A. in Math, minor in Biology, 2015

Interests:

Statistical computing and data science
Genomics and statistical genetics

Relevant Skills

Programming Languages: R, Shiny, Python, C++, Unix

Statistical genetics and genomics software: PLINK, VCFtools, BEAGLE

Lab and Research Experience

Johns Hopkins University, Biostatistics - Baltimore, MD 9/2017- present
- Topic: Rare variant association testing in case-parent trio data
- Advisors: Prof. Margaret Taub, Prof. Terri Beaty
- Created R workflow for detecting rare variants associated with disease risk in case-parent trios
- Wrote literature review of rare-variant analysis methods for case-parent trio study designs, including extensions of transmission disequilibrium tests, conditional logistic regression for matched case-controls, and scan statistic
- Analyzed cleft palate case-parent trio whole-genome sequencing data

Johns Hopkins University, High-Throughput Biology Center - Baltimore, MD 5/2014- 8/2014
- Topic: A search engine for proteins and researchers
- Advisor: Prof. Joel S. Bader
- Implementing a search engine to identify highly-relevant scientists that study a given protein
- Utilized NCBI e-Utilities to download large datasets and implemented PageRank search algorithm

Swarthmore College, Department of Mathematics and Statistics - Swarthmore, PA 5/2013- 8/2013
- Topic: Estimating the tempo of the Cambrian Explosion
- Advisor: Prof. Steve C. Wang
- Investigated accuracy of the Cambrian fossil record using Monte Carlo simulation algorithm
- Wrote functions to account for inherent margins of error in radiometric dating and linearly-changing sedimentation rates in R

University of California-Berkeley, Molecular and Cellular Biology - Berkeley, CA 6/2012- 8/2012

- Topic: Genetics of tooth number determination in threespine stickleback
- Advisors: Nick Ellis, Prof. Craig T. Miller
- Identified tooth gain QTL in lab stock of high-toothed population of Paxton Lake benthic fish
- Performed in situ hybridization to determine *ap2* and *bmp6* expression patterns

Course Projects:

Advanced Data Science II project 10/2017-12/2017

- Title: Police Violence Visualization
- Co-author: Kenneth Morales
- Created an interactive heat map of lethal police encounters in the U.S. using R Shiny and Leaflet

Advanced Data Science I project 8/2016-10/2016

- Title: Detecting suicidal ideation in Reddit text posts
- Co-author: Lacey Etzkorn
- Used Python scraper to download text posts from Reddit's depression message board
- Created logistic regression model in R to predict presence of suicidal ideation

Awards

NSF Evo-Devo-Eco Network Undergraduate Training Grant	2012
Sigma Xi Grant-in-Aid of Research - $600	2012

Publications

Cleves PA, Hart JC, Agoglia RM, Jimenez MT, Erickson PA, **Gai L**, et al. (2018) An intronic enhancer of Bmp6 underlies evolved tooth gain in sticklebacks. PLoS Genet 14(6): e1007449. https:// doi.org/ 10.1371/journal.pgen.1007449

J.C. Talbot, M.B. Walker, T.J. Carney, T.R. Huycke, Y.L. Yan, R.A. Bremiller, **L. Gai**, A. Delaurier, J.H. Postlethwait, M. Hammerschmidt et al. *fras1* shapes endodermal pouch 1 and stabilizes zebrafish pharyngeal skeletal development. Development, 139 (2012), pp. 2804–2813

Presentations

L. Gai, N.A. Ellis, C.T. Miller. "Genetics of evolved tooth number divergence in threespine stickleback." Invited speaker at the *Society for Developmental Biology*. Snowbird, UT (July 2015).

C. Wang, **L. Gai**, S.C. Wang, J.L. Moore, S.M. Porter, A.C. Maloof. "Estimating the tempo of the Cambrian explosion." Presented at the *Geological Society of America.* Denver, CO (October 2013).

L. Gai, N.A. Ellis, C.T. Miller. "Genetics of evolved tooth number divergence in threespine stickleback." Presented at the *Society for Integrative and Comparative Biology*. San Francisco, CA (January 2012).

Coursework

Graduate coursework:

Statistics:

Ph.D core curriculum (year-long sequence):
Statistical Theory, Statistical Methods, Probability Theory, Advanced Data Science I-II

Electives:
Advanced Statistical Computing, Statistical Machine Learning,
Survival Analysis I , Longitudinal Data Analysis, Multilevel Modeling (audit)

Biology: Genetic Epidemiology II and IV, Genomics for Public Health, Epidemiology I

Undergraduate coursework:

Mathematics: Linear Algebra, Multivariable Calculus, Differential Equations, Real Analysis I
Computer science: Data Structures and Algorithms in C++, Bioinformatics in Python
Biology: Mathematical Modeling in Biology (directed reading), Developmental Biology

Teaching Experience

TA:

Statistics for Laboratory Scientists III-IV	2019
Statistics for Laboratory Scientists I-II	2017, 2018
Statistical Methods for Public Health I-IV	2017, 2019
Statistical Reasoning in Public Health I-II	2016, 2018
Data Analysis Workshop I, Summer Institute	2016

Service

Biostatistics Journal Club co-coordinator, 2017-2018

Volunteered at:

Institute of Mathematical Statistics New Researchers Conference, July 2017
Course Monitor for Bayesian Time Series, JSM 2017